przegląd krytyczny
critical review

Epistemological aspect of topic modelling in the social sciences: Latent Dirichlet Allocation

PRZEGLĄD KRYTYCZNY
2022 / 4(1): 7–16
ISSN: 2657–8964
DOI: 10.14746/pk.2022.4.1.1

Mariusz Baranowski[1]

[1] Adam Mickiewicz University, Faculty of Sociology, Szamarzewskiego 89C, 60-568 Poznań, Poland. ORCID: 0000-0001-6755-9368, Email: mariusz.baranowski@amu.edu.pl

ABSTRACT: Aware of the challenges faced by the social sciences in publishing a massive volume of research papers, it is worth looking at a novel but no longer so new ways of machine learning for the purposes of literature review. To this end, I explore a probabilistic topic model called Latent Dirichlet Allocation (LDA) in the context of the epistemological challenge of analysing texts on social welfare. This paper aims to describe how the LDA algorithm works for large corpora of data, along with its advantages and disadvantages. This preliminary characterisation of an inductive method for automated text analysis is intended to give a brief overview of how LDA can be used in the social sciences.
KEYWORDS: Latent Dirichlet Allocation (LDA), topic modelling, social sciences, social welfare, automated text analysis

INTRODUCTION

The mass "production" of scientific papers in the form of books, articles, chapters in collective monographs, research or post-conference reports represents, on the one hand, a positive effect of the researchers' communication, but on the other hand, a severe challenge in reviewing and selecting their content. From an epistemological perspective, which deals with ways of knowing the world and studying knowledge about it, the inability to reliably review research findings is highly problematic. "The theory of knowledge and justification", as epistemology is also referred to as such (Audi, 2003), needs to develop ways of meta-analysing the content of the products of scientific research, as well as methods of evaluating them. And since we live in a technologically networked society (Baranowski, 2021), in which mobile applications dominate, a systematic literature review must also include these digital footprints

(Genc-Nayebi & Abran, 2017).

Since the review of research in a field is essential for the correct preparation of the research questions of a new project or the interpretation of results already obtained, therefore this "process is usually performed manually, which means that reviewers need to read thousands of citations during the screening phase, due to the rapid growth of the (...) literature, making it an expensive and time-consuming process" (Mo, Kontonatsios, & Ananiadou, 2015). Additionally, the influence of those reviewing the literature (reviewer bias) is another constraint, forcing the search for ways to overcome these limitations, hence "methods such as machine learning, text mining (Ananiadou, Rea, Okazaki, Procter, & Thomas, 2009; O'Mara-Eves, Thomas, McNaught, Miwa, & Ananiadou, 2015), text classification (García Adeva, Pikatza Atxa, Ubeda Carrillo, & Ansuategi Zengotitabengoa, 2014) and active learning (Wallace, Small, Brodley, & Trikalinos, 2010; Wallace, Trikalinos, Lau, Brodley, & Schmid, 2010) have been used to partially automate this process, in order to reduce the workload, without sacrificing the quality of the reviews" (Mo et al., 2015).

BACKGROUND

Let us take examples of publications on "social welfare" and "welfare state" exclusively from the Scopus database. Figure 1 below shows the number of papers published each year from 1948 to 2020 containing the words "social welfare" and "welfare state" in the texts' titles, abstracts or keywords. Not only did the number of published texts start to increase rapidly at the end of the 1990s, but particularly in the context of social welfare, we are dealing with both a phenomenon addressed and defined differently by different disciplines (Baranowski, 2017, 2019, 2022a, 2022b; Forder, Caslin, Ponton, & Walklate, 2018).

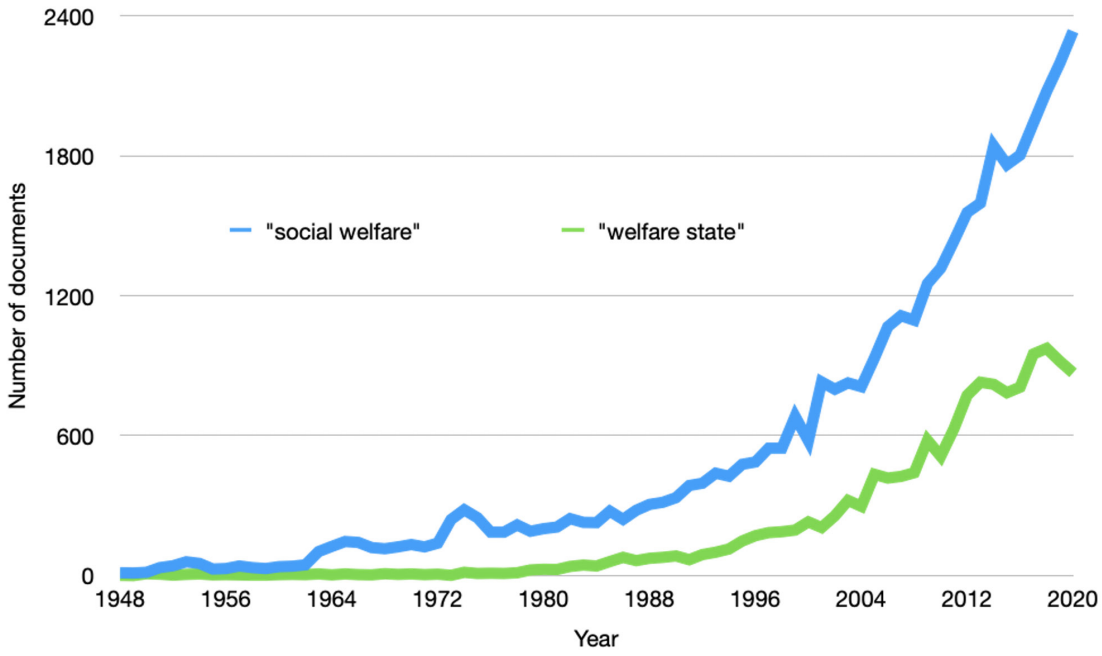


Figure 1. "Social welfare" versus "welfare state"
Source: own elaboration based on Scopus.

Figure 2 provides information on the attribution of texts in which the terms "social welfare" or "welfare state" appear to particular subject areas. It turns out that while the welfare state is dominated by journals assigned to the social sciences (56 per cent), the situation is quite different regarding social welfare. In the latter case, the most significant number of publications were classified under medicine (34 per cent) and less than 25 per cent under social sciences. It is worth noting that Economics, Econometrics and Finance delineate a different subject area in the Scopus classification.

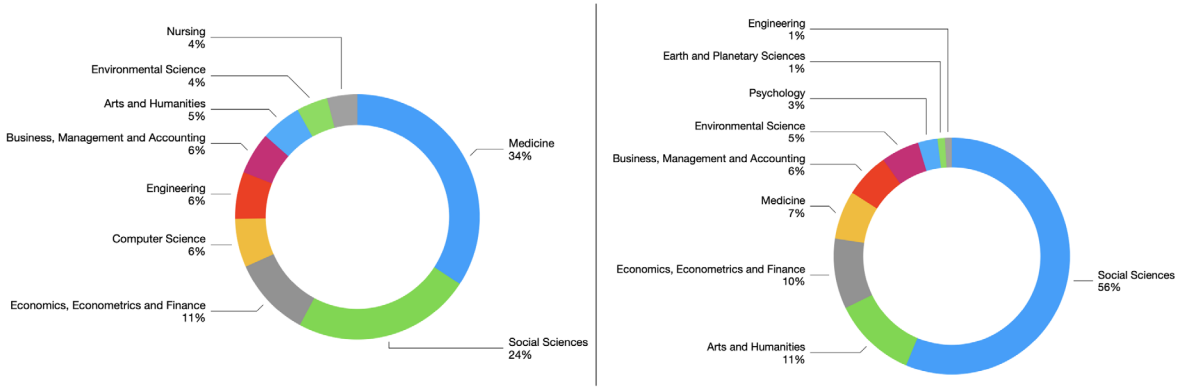


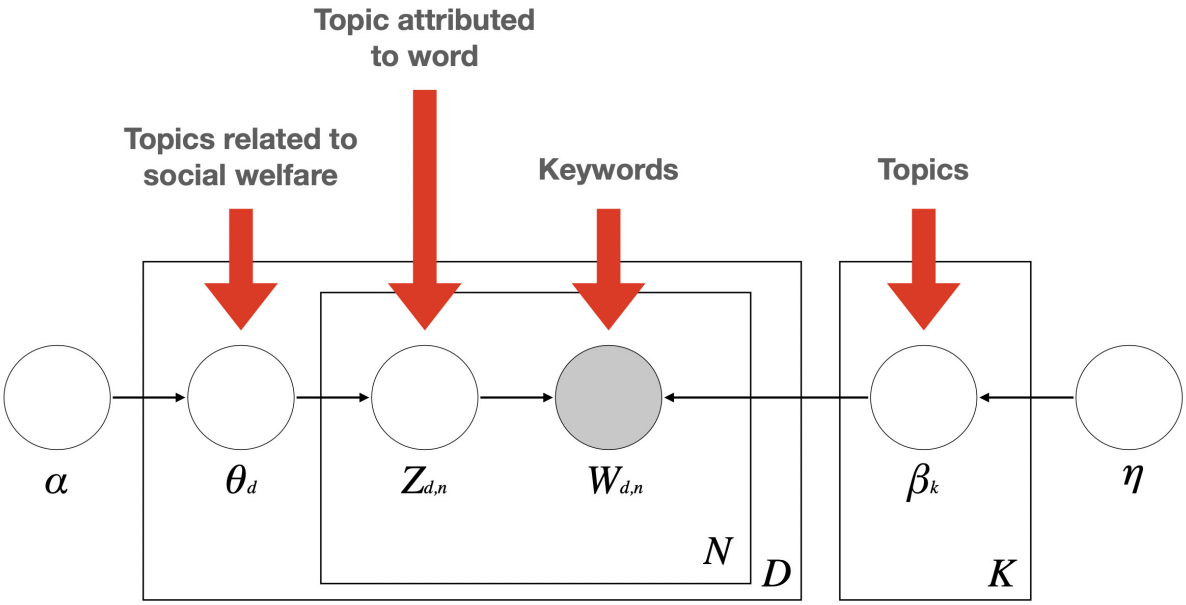
Figure 2. Subject area for 'social welfare' (on the left) and 'welfare state' (on the right)
Source: own elaboration based on Scopus.

Given both the number of publications and the different thematic fields on social welfare, it cannot be ruled out that journals classified under Environmental Science or Business, Management and Accounting should not be included in approaches from, for example, sociology (Baranowski & Cichocki, 2021). This requires a lot of resources to conduct a literature review with researchers or advanced methods to analyse large data corpora. And this is where we come to LDA topic modelling.

DESCRIPTION OF THE METHOD

Let's start with the fact that topic modelling is a "statistical method which aims to discover an abstract topic in a set of documents" (Pandur, Dobša, & Kronegger, 2020), and additionally, "it is an unsupervised machine learning technique because it does not require a training dataset or a predefined documents based on words or expressions with similar meaning" (Pandur et al., 2020). One of the most commonly used methods of this type is Latent Dirichlet Allocation (LDA), first introduced by Blei, Ng, and Jordan (2003). Importantly, LDA is actually a mixture of models because "documents are considered as a mixture of topics (Blei et al., 2003; Thomas Hofmann, 1999; Thomas Hofmann, 2001; Steyvers & Griffiths, 2007). They are also known as admixture because its segments are itself mixture of other segments (Heinrich, 2009)" (Chauhan & Shah, 2021).

LDA as a generative probabilistic topic model serves to "uncover latent semantic structures from a set of documents, *D*. LDA models documents as discrete distributions over *K* latent topics, and every topic is modeled as a discrete distribution over the fixed vocabulary" (Syed & Spruit, 2018, p. 195). The latent semantic structure shown in Figure 3 expressed in terms of topics (β) is most salient about LDA yet elusive in manual coding. Applied to the social welfare data, θ_d stands for topic proportions. However, let us remember that „β, θ, and Z are unobserved, and the goal is to determine them from the observed variables (i.e. the words within the documents)" (Syed & Spruit, 2018).



α – Dirichlet priority parameter (distribution per document/topic)
θ_d – proportion of topics in the article *d*
$Z_{d,n}$ – assigned topic to *n*-word in article *d*
$W_{d,n}$ – the observed word *n* in the article *d*
β_k – word distribution
η – Dirichlet priority parameter (distribution per topic/word)

Figure 3. Graphical representation of LDA topic modelling
Source: own elaboration based on Chauhan & Shah (2021), Mo et al. (2015), Pandur et al. (2020), Syed & Spruit (2018).

The entire generative process of determining themes (topics) follows the following pattern (Syed & Spruit, 2018):

1) For every topic $k = \{1,\ldots, K\}$

a) draw a distribution over the vocabulary *V*, $\beta_k \sim \text{Dir}(\eta)$

2) For every document *d*

a) draw a distribution over topics, $\theta_d \sim \text{Dir}(\alpha)$ (i.e. per-document topic proportion)

b) for each word *w* within document *d*

i) draw a topic assignment, $z_{d,n} \sim \text{Mult}(\theta_d)$, where $z_{d,n} \in \{1,\ldots, K\}$ (i.e. per-word topic

assignment)

ii) draw a word $w_{d,n}$ ~ Mult($\beta z_{d,n}$), where $w_{d,n} \epsilon \{1,..., V\}$

Based on these guidelines, the distribution of latent and observable variables takes the form of the following equation:

$$p(\beta_K, \theta_D, z_D, w_D|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_K|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{d,k})$$

This way, from substantial text datasets, we can generate a certain number of central topics with associated authors, journals and metadata on, for example, the number of citations of each document.

DISCUSSION

Text mining methods such as LDA topic modelling offer researchers systematic content analysis for large data sets for such "a soft, wide and complex subject domain as Welfare" (Wormell, 2000, p. 203). However, the palette of available analysis tools originally developed in computer science, advanced statistics and computational linguistics is much broader (Figure 4).

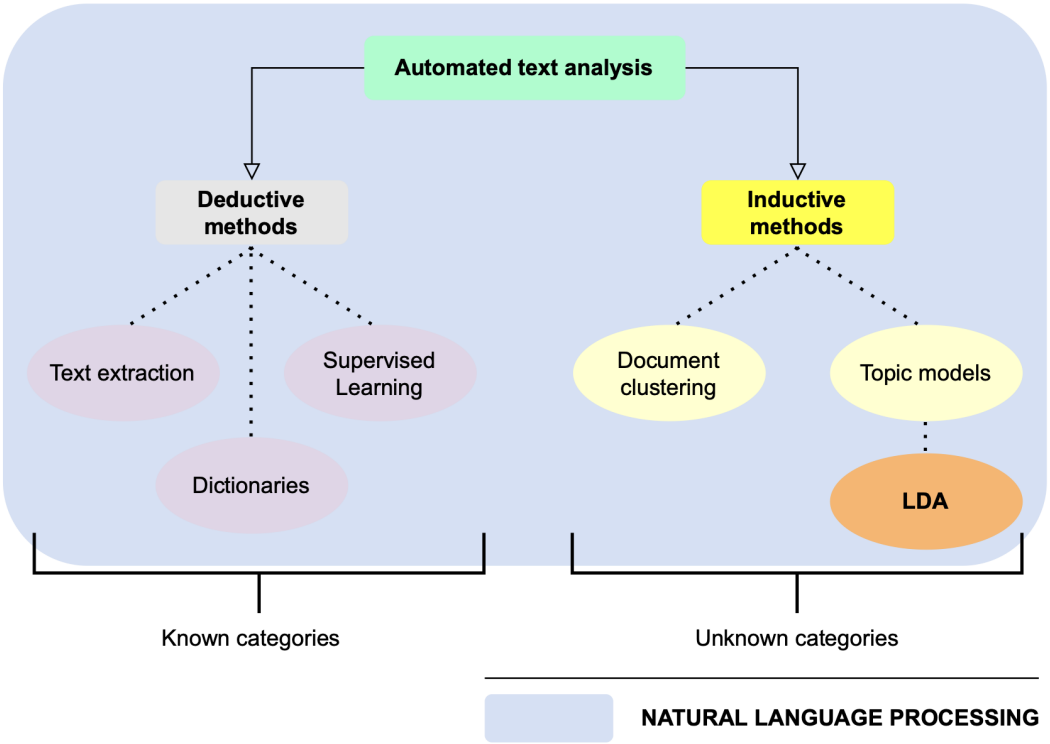


Figure 4. Types of automated text analysis
Source: own elaboration based on Günther and Quandt (2016, p. 76).

First of all, as can be seen from the figure above, automated text analysis has different methods that offer different solutions and require different workloads from researchers. Thus, deductive text analysis methods require the researcher to define the main categories manually even before the actual study begins. This requires time and money but, on the other hand, provides solutions for advanced content research. As the German researchers note, "sentiment analyses are well-known examples of this approach, evaluating the occurrence of positive and negative terms in a document based on a specific list (=dictionary)" (Günther & Quandt, 2016, p. 77).

More relevant from the perspective of the purpose of this article are inductive methods, in particular LDA, or "a statistical model of language" (Mohr & Bogdanov, 2013, p. 547). Well, this type of method is "especially useful when a researcher has little knowledge on the contents of a document collection. When crafting a codebook for a manual content analysis, setting up a dictionary, or specifying rules to extract text features are not feasible, fully automated approaches are a valuable aid to get to know more about the documents' contents" (Günther & Quandt, 2016, p. 77).

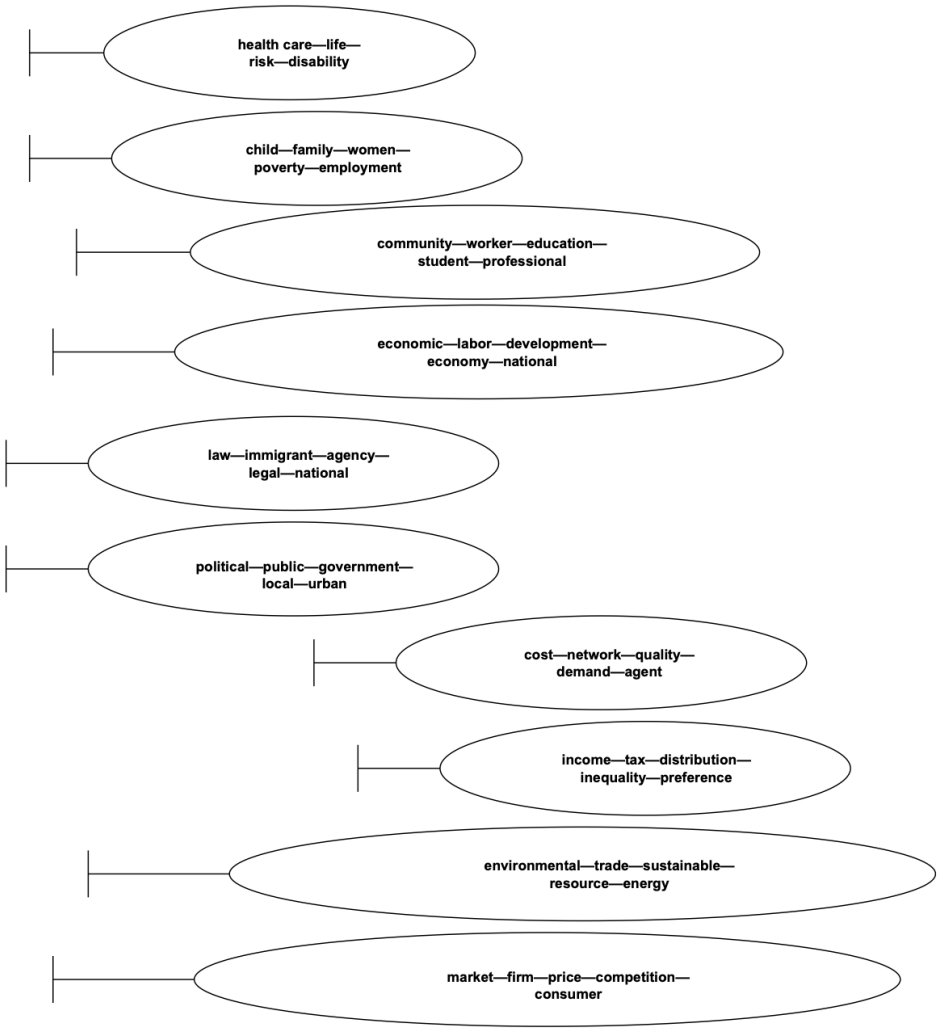


Figure 5. Topic modelling applied to social welfare on a trial basis
Source: own elaboration.

Let's look at ten training-generated topics in analysing tens of thousands of abstracts of documents indexed in the Scopus database. Generating these topics required several preparatory steps, such as data cleaning, modelling and parameter setting, because "LDA-techniques belong to the general approaches known as 'bag-of-words', which amounts to treating every document as an unordered list of tokens (usually but not necessarily individual words or common phrases)" (Baranowski & Cichocki, 2021, p. 11). All this to obtain—as Evans and Aceves (2016, p. 23) put it—"a wealth of sociologically relevant data".

The ten topics generated can be interpreted in light of additional data, such as journal type, main discipline, authors' affiliation, citations, or funding institutions. A secondary cluster analysis can also be performed to group the themes. In Figure 5, two themes, i.e. (1) *health—care—life—risk—disability* and (2) *child—family—women—poverty—employment*, are closely related, while others show weaker coherence.

Those working with social welfare issues can use LDA to map an entire thematic area or explore a particular part of it. It all depends on the objectives and research strategy (McFarland et al., 2013; Pääkkönen & Ylikoski, 2021).

CONCLUSIONS

Thomas S. Kuhn, in *The Structure of Scientific Revolutions* [1962] (1970, p. 10) defined "normal science" as "research firmly based upon one or more past scientific achievements, achievements that some particular scientific community acknowledges for a time as supplying the foundation for its further practice". However, for real progress in science, a radical change is also needed in these "past" achievements, which, although they have contributed to the expansion of knowledge in a particular field, at a certain point become an inadequate perspective for the study of reality.

Given, as Fiona Williams, Jennie Popay and Ann Oakley (2014, p. 2) put it, "a new framework for social research in the welfare field", it is worth reflecting on ways of exploring the phenomenon of "social welfare" in ways that deviate from accepted patterns of practice. And this is by no means about the political-economic context of the determinants of social welfare, which—there is no doubt about it—still determines the critical determinants of institutional varieties of state interventions. This article is not only part of a critique of the narrow (especially prevalent in the USA) understanding of the concept of welfare, about which Joe R. Feagin (1973, p. 321) wrote almost five decades ago that "what many politicians and commentators have in mind when they speak of 'welfare' are (…) public assistance programs" (cf. Hadley & Hatch, 2019; Plant, 2019). This article aimed to highlight the richness of the themes behind social welfare using LDA topic modelling.

FUNDING: This work was supported by the National Science Centre, Poland, under research project "Social welfare in the light of topic modelling: A preliminary study", no 2021/05/X/HS6/00067.

REFERENCES

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting Systematic Reviews Using Text Mining. *Social Science Computer Review, 27*(4), 509-523. doi:10.1177/0894439309332293

Audi, R. (2003). *Epistemology: A contemporary introduction to the theory of knowledge* (Second ed.). New York and London: Routledge.

Baranowski, M. (2017). Welfare sociology in our times. How social, political, and economic uncertainties shape contemporary societies. *Przegląd Socjologiczny, 66*(4), 9-26. doi:https://doi.org/10.26485/PS/2017/66.4/1

Baranowski, M. (2019). The Struggle for social welfare: towards an emerging welfare sociology. *Society Register, 3*(2), 7-19. doi:https://doi.org/10.14746/sr.2019.3.2.01

Baranowski, M. (2021). The Sharing Economy: Social Welfare in a Technologically Networked Economy. *Bulletin of Science, Technology & Society, 41*(1), 20-30. doi:10.1177/02704676211010723

Baranowski, M. (2022a). Myths, Narratives and Welfare States: The Impact of Stories on Welfare State Development. *Contemporary Sociology, 51*(3), 202-204. doi:10.1177/00943061221090769h

Baranowski, M. (2022b). Radicalising Cultures of Uneven Data-Driven Political Communication. *Knowledge Cultures, 10*(2), 145-155. doi: https://doi.org/10.22381/kc10220227

Baranowski, M., & Cichocki, P. (2021). Good and bad sociology: Does topic modelling make a difference? *Society Register, 5*(4), 7-22. doi:https://doi.org/10.14746/sr.2021.5.4.01

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022. doi:https://dl.acm.org/doi/10.5555/944919.944937

Chauhan, U., & Shah, A. (2021). Topic Modeling Using Latent Dirichlet allocation: A Survey. *ACM Computing Surveys, 54*(7), Article 145. doi:10.1145/3462478

Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual review of sociology, 42*(1), 21-50. doi:10.1146/annurev-soc-081715-074206

Feagin, J. R. (1973). Issues in welfare research: A critical overview. *Social Science Quarterly, 54*(2), 321-328. Retrieved from https://www.jstor.org/stable/42859163

Forder, A., Caslin, T., Ponton, G., & Walklate, S. (2018). *Theories of welfare*. London: Routledge.

García Adeva, J. J., Pikatza Atxa, J. M., Ubeda Carrillo, M., & Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications, 41*(4), 1498-1508. doi:https://doi.org/10.1016/j.eswa.2013.08.047

Genc-Nayebi, N., & Abran, A. (2017). A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software, 125*, 207-219. doi:https://doi.org/10.1016/j.jss.2016.11.027

Günther, E., & Quandt, T. (2016). Word Counts and Topic Models. *Digital Journalism,*

4(1), 75-88. doi:10.1080/21670811.2015.1093270

Hadley, R., & Hatch, S. (2019). *Social Welfare and the Failure of the State: Centralised Social Services and Participatory Alternatives*. Abingdon and New York: Routledge.

Heinrich, G. (2009). *Parameter estimation for text analysis*. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.6555&rep=rep1&type=pdf

Hofmann, T. (1999). *Probabilistic latent semantic analysis*. Paper presented at the Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Stockholm, Sweden.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning, 42*(1), 177-196. doi:10.1023/A:1007617005950

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (Second ed.). Chicago: The University of Chicago Press.

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics, 41*(6), 607-625. doi:https://doi.org/10.1016/j.poetic.2013.06.004

Mo, Y., Kontonatsios, G., & Ananiadou, S. (2015). Supporting systematic reviews using LDA-based document representations. *Systematic Reviews, 4*, 172. doi:10.1186/s13643-015-0117-0

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics, 41*(6), 545-569. doi:https://doi.org/10.1016/j.poetic.2013.10.001

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews, 4*(1), 5. doi:10.1186/2046-4053-4-5

Pääkkönen, J., & Ylikoski, P. (2021). Humanistic interpretation and machine learning. *Synthese, 199*(1), 1461-1497. doi:10.1007/s11229-020-02806-w

Pandur, M. B., Dobša, J., & Kronegger, L. (2020). *Topic Modelling in Social Sciences: Case Study of Web of Science*. Paper presented at the Central European Conference on Information and Intelligent Systems, Varazdin.

Plant, R. (2019). Needs and welfare. In N. Timms (Ed.), *Social welfare: Why and How?* (pp. 103-122). Abingdon and New York: Routledge.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In *Handbook of latent semantic analysis*. (pp. 427-448). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Syed, S., & Spruit, M. (2018, 31 Jan.-2 Feb. 2018). *Selecting Priors for Latent Dirichlet Allocation*. Paper presented at the 2018 IEEE 12th International Conference on Semantic Computing (ICSC).

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2010). *Active learning for biomedical citation screening*. Paper presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA. https://doi.org/10.1145/1835804.1835829

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics, 11*(1), 55. doi:10.1186/1471-2105-11-55

Williams, F., Popay, J., & Oakley, A. (2014). Changing Paradigms of Welfare. In F. Williams, J. Popay, & A. Oakley (Eds.), *Welfare Research: A Critical Review* (pp. 2-17). London and New York: Routledge.

Wormell, I. (2000). Bibliometric Analysis of the Welfare Topic. *Scientometrics, 48*(2), 203-236. doi:10.1023/A:1005696722014